Investigation of mathematical model of passenger preferences

Cite as: AIP Conference Proceedings **2172**, 080001 (2019); https://doi.org/10.1063/1.5133559
Published Online: 13 November 2019

Galina Timofeeva

Investigation of Mathematical Model of Passenger Preferences

Galina Timofeeva[1,2,a)]

[1]*Ural State University of Railway Transport, Kolmogorov str., 66, Ekaterinburg, 620034, Russia*
[2]*Ural Federal University named after First President of Russia B. N. Yeltsin, Mira str., 19, Ekaterinburg, 620002, Russia*

[a)]Corresponding author: Gtimofeeva@usurt.ru

Abstract. A mathematical model for describing passenger preferences is studied. The model was proposed earlier by the authors for predicting passenger traffic for high-speed links of the railway network. Choosing a route by a randomly selected passenger is considered as a problem of minimizing of the generalized transportation cost. This cost is a linear combination of two criteria (a cost and time of the transportation) and depends on a random parameter of passenger preferences. The properties of the model are investigated. The estimation problem for the random parameter distribution via statistical data on the passengers preferences is studied.

PROBABILISTIC MODEL FOR PASSENGER PREFERENCE

We study the probabilistic model of choosing the transport type proposed in [1, 2]. Suppose a consumer (a passenger) had a choice between n possible alternatives $e_1, \ldots, e_n$ before the transport network structure was changed. The set of these alternatives is denoted by $A_0 = \{e_1, \ldots\ldots, e_n\}$, where $\{e_i, i = \overline{1, n}\}$ is the basis in the space R^n. Let a new mode of transport (a new route) $\hat{e}_{n+1}$ be added. Denote by A_1 the set of alternatives after the changing of the network structure.

Preferences of a randomly selected passenger are described by a random vector (ξ_0, ξ_1), where ξ_i are discrete random variables that depend on each other and take the values in A_i, $i = 0, 1$. The random value ξ_0 reflects of the distribution of preferences before the introduction of a new link in the transport network and ξ_1 after the introduction.

Let denote the probability distribution of ξ_0 and ξ_1 by $q^{(0)}$ and $q^{(1)}$ respectively, where $q^{(0)} = (q_1^{(0)}, \ldots, q_n^{(0)})$, $q^{(1)} = (q_1^{(1)}, \ldots, q_{n+1}^{(1)})$ and

$$q_i^{(0)} = Pr\{\xi_0 = e_i\}, \quad i = \overline{0, n}. \tag{1}$$

$$q_i^{(1)} = Pr\{\xi_1 = \hat{e}_i\}, \quad i = \overline{0, n+1}, \tag{2}$$

here $\{\hat{e}_1, \ldots, \hat{e}_{n+1}\}$ is the basis of the space R^{n+1}.

Bicriterial Approach

Let us denote by c_i a trip cost and by t_i a trip time, $i = \overline{1, n}$. We assume that there are no coincident parameter values, i.e. the points $(c_i, t_i) \in R^2$, $i = \overline{1, n}$ are different.

In our consideration other conditions of travel (such as the convenience of timetables, etc.) are not taken into account, so we obtain the optimization problem with two criteria for describing the preferences of one passenger between alternatives $\{e_1, \ldots, e_n\}$:

$$\begin{gathered} x_1 + \ldots + x_n = 1, \quad x_i \in \{0; 1\}, \quad i = \overline{1, n}, \\ T(X) = t_1 x_1 + \ldots + t_n x_n \to \min_X \\ C(X) = c_1 x_1 + \ldots + c_n x_n \to \min_X . \end{gathered} \tag{3}$$

Proceedings of the 45th International Conference on Application of Mathematics in Engineering and Economics (AMEE'19)
AIP Conf. Proc. 2172, 080001-1–080001-7; https://doi.org/10.1063/1.5133559
Published by AIP Publishing. 978-0-7354-1919-3/$30.00

Here the choice of the i-th route is written as $X = (x_1, \dots, x_n) = e_i \in R^n$.

Definition 1. The solution $X^{(1)}$ is preferable to the solution (dominates over the solution) $X^{(2)}$, if one of two conditions holds:

$$T(X^{(1)}) \leq T(X^{(2)}) \wedge C(X^{(1)}) < C(X^{(2)})$$

or

$$T(X^{(1)}) < T(X^{(2)}) \wedge C(X^{(1)}) \leq C(X^{(2)}).$$

The set of non-dominant (Pareto-optimal) solutions to the bicriterial problem (3) is denoted by $E_0 \subseteq A_0$. To describe a passenger preference we will use the "generalized trip cost" introduced in papers [3, 4]. It is a linear function of the trip cost and the trip time

$$f(X) = C(X) + \tau T(X),$$

where $\tau > 0$ may be considered as time unit price.

The problem of "generalized trip cost" minimization has the form:

$$\begin{array}{l} x_1 + \ldots + x_n = 1, \quad x_i \in \{0; 1\}, \quad i = \overline{1, n}, \\ f_\tau(X) = C(X) + \tau T(X) \to \min\limits_X . \end{array} \tag{4}$$

The set of optimal solutions to parametrical optimization problem (4) is denoted by $X(\tau)$, where $\tau \geq 0$.

PROPERTIES OF OPTIMIZATION PROBLEM SOLUTIONS

Let us consider properties of the solution $X(\tau)$ of parametrical optimization problem (4). Denote by $\mathbf{X}_A$ the union of solution sets for all values of the parameter

$$\mathbf{X}_A = \bigcup_{\tau > 0} X(\tau).$$

Property 1. The set of parametrical optimization problem solutions is a subset of Pareto-optimal solutions set:

$$\mathbf{X}_A \subset E_0,$$

where E_0 is the set of Pareto-optimal solutions to bicriterial problem (4).

Proof. Let the alternative $e_j \notin E_0$, i.e. it is dominated and there exists $e_m \in A_0$ such that $c_j \geq c_m$ and $t_j \geq t_m$, and at least one inequality is strict. From

$$f_\tau(e_j) = c_j + \tau t_j \quad \forall j = \overline{1, n}$$

follows $f_\tau(e_j) > f_\tau(e_m)$ for all $\tau > 0$. Then the set of the solution to the optimization problem $X(\tau)$ doesn't contain e_j for any $\tau > 0$, i.e. $e_j \notin \mathbf{X}_A$.

The proven statement for arbitrary $e_j \in A_0$

$$(e_j \notin E_0) \Rightarrow (e_j \notin \mathbf{X}_A)$$

results in

$$(e_j \in \mathbf{X}_A) \Rightarrow (e_j \in E_0)$$

and $\mathbf{X}_A \subset E_0$.

Note that not all elements of the set of all effective solutions E_0 necessarily are solutions to the parametric problem, i.e. the inclusion

$$\mathbf{X}_A \subset E_0$$

may be strict.

Example 1. Consider a problem of choosing one of three alternatives $A_0 = \{e_1, e_2, e_3\}$. Let values of the first and the second criteria (trip price and trip time) are respectively (1; 5), (4; 4) and (5; 1) (see Fig.1).

For this data the set of Pareto-optimal solutions to (3) coincides with the set of alternatives, i.e. $E_0 = A_0$, since all of them are not dominated. The solution to the parametrical optimization problem (4) has the form:

$$X(\tau) = \begin{cases} e_1 \text{ if } \tau < 1 \\ e_3 \text{ if } \tau > 1 \\ e_1 \cup e_3 \text{ if } \tau = 1. \end{cases}$$

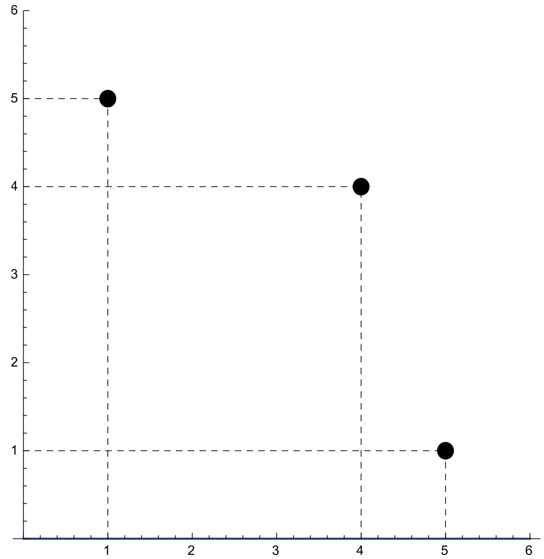

FIGURE 1. Criteria values (trip price and trip time) for three alternatives.

Thus, $\mathbf{X}_A = e_1 \cup e_3$ and $e_2 \notin \mathbf{X}_A$, i.e. the inclusion $\mathbf{X}_A \subset E_0$ is strict.

Property 2. $e_k \in X(\tau)$ if and olny if $e_k \in E_0$ and

$$L_0(k) \leq \tau \leq R_0(k), \tag{5}$$

where

$$L_0(k) = \max_{t_j > t_k, j=\overline{1,n}} \left\{0, \frac{c_k - c_j}{t_j - t_k}\right\}, \quad R_0(k) = \min_{t_j < t_k, j=\overline{1,n}} \left\{\frac{c_k - c_j}{t_j - t_k}, +\infty\right\}. \tag{6}$$

Proof. Let $e_k \in E_0$, i.e. e_k is a Pareto-optimal alternative of bicriterial problem (3). The condition $e_k \in X_0(\tau)$ is equivalent

$$f_\tau(e_k) \leq f_\tau(e_j) \ \ \forall j = \overline{1,n}. \tag{7}$$

Taking into account

$$f_\tau(e_j) = c_j + \tau t_j\,, \quad j = \overline{1,n},$$

we get that relation (7) is equivalent to the fulfillment of the conditions for all $j = \overline{1,n}$:

$$\tau \geq \frac{c_k - c_j}{t_j - t_k} \ \text{ if } \ t_j > t_k$$

$$\tau \leq \frac{c_j - c_k}{t_k - t_j} \ \text{ if } \ t_j < t_k.$$

It should be mention that $c_j < c_k$ if $t_j > t_k$ for all $e_j, e_k \in E_0$.

If $t_j \leq t_k$ for all $j = \overline{1,n}$, then $L_0(k) = 0$. If $t_j \geq t_k$ for all $j = \overline{1,n}$, then $R_0(k) = +\infty$.

Thus, $e_k \in X_(\tau)$ if and only if conditions (5), (6) hold.

Corollary 1. If $e_k \in E_0$ and

$$L_0(k) < \tau < R_0(k), \tag{8}$$

where $L_0(k)$ and $R_0(k)$ are defined by equations (6), then $X(\tau) = e_k$.

Corollary 1 follows from the proof of Property 2.

Remark. Some of the intervals $[L_0(k); R_0(k)]$ may be empty, it means that $e_k \notin \mathbf{X}_A$. For Example 1 $L_0(2) = 3$, $R_0(2) = 1/3$, inequality (8) is not true for $k = 2$ and $e_2 \notin X_A$.

Corollary 2. The set of all solution to parametric optimization problem (4) $X(\tau)$ consists of more then one point for $\tau = L_0(k)$ if $L_0(k) > 0$ and for $\tau = R_0(k) < +\infty$.

Proof. Let denote by J a number of element at which the maximum

$$\max_{t_j > t_k, j=\overline{1,n}} \left\{ 0, \frac{c_k - c_j}{t_j - t_k} \right\} = L_0(k) > 0$$

is reached, i.e.

$$\frac{c_k - c_J}{t_J - t_k} = \max_{t_j > t_k, j=\overline{1,n}} \left\{ \frac{c_k - c_j}{t_j - t_k} \right\} = L_0(k).$$

So $c_k - c_J = \tau(t_J - t_k)$ if $\tau = L_0(k)$, therefore $c_k + \tau t_k = c_J + \tau t_J$ and $f_\tau(e_J) = f_\tau(e_k)$.

From Property 2 follows $e_k \in X_0(L_0(k))$. Thus we get

$$e_k \cup e_J \in X_0(\tau) \text{ if } \tau = L_0(k) > 0$$

and $X_0(L_0(k))$ consists of more than one point.

The statement about $X(\tau)$ for $\tau = R_0(k)$ is proved similarly.

RANDOM PREFERENCE

If the parameter τ (a value of time unit) is the same for all passengers, then they choose the same solution to the parametrical optimization problem (4), i.e. one transport mode. However, it does not happen, so we assume that the value of time costs is different for different passengers and they choose different solutions. The model of the random passenger choice as a solution to the optimization problem with a random objective function was proposed in [1].

Let the random variable $\theta = \theta(\omega)$ reflect the preference of a randomly chosen passenger ω. His choice of the transport mode before the introduction of a new link is denoted by $X_0(\theta)$.

It is proposed that $X_0(\theta)$ is a solution to the linear optimiation problem (4) with a random parameter $\tau = \theta(\omega)$.

Suppose it become possible to use an additional type of transport on the existing arc of the transport network (for example, high-speed transport), so a set of alternatives expands and becomes equal to A_1, where

$$A_1 = \{\hat{e}_1, \ldots, \hat{e}_{n+1}\},\ \hat{e}_i = (e_i, 0) \in R^{n+1},\ i = \overline{1,n},\ \hat{e}_{n+1} = (0, \ldots, 0, 1).$$

The choice problem is written in the form of a parametrical optimization problem, similar to problem (4):

$$\begin{aligned} &x_1 + \ldots + x_{n+1} = 1, \quad x_i \in \{0; 1\}, \quad i = \overline{1, n+1}, \\ &f_\tau^{(1)}(X) = C_1(X) + \tau T_1(X) \to \min_X, \quad X \in R^{n+1}, \end{aligned} \tag{9}$$

where

$$\begin{aligned} C_1(X) &= c_1 x_1 + \ldots + c_n x_n + c_{n+1} x_{n+1}, \\ T_1(X) &= t_1 x_1 + \ldots + t_n x_n + t_{n+1} x_{n+1}. \end{aligned} \tag{10}$$

Lemma 1. Let $X_0(\tau) = e_i$ for a fixed $\tau > 0$, then a set $X_1(\tau)$ of solutions to optimization problem (9)–(10) satisfies

$$X_1(\tau) \subseteq \hat{e}_i \cup \hat{e}_{n+1}.$$

This statement follows from the form of optimization problems (4) and (9)–(10).

Transition Probabilities

Let $\xi_1(\theta)$ denote a choice of the passenger after the introduction of the new link. It is proposed that $\xi_1(\theta)$ is the solution to the optimization problem (9)–(10) with the same random parameter $\tau = \theta(\omega)$.

The values $\xi_0(\theta)$ and $\xi_1(\theta)$ are discrete random values related to each other. Let's calculate their joint distribution and transition probabilities p_{ij}, where θ is the same continuous random valuable at both stages (i.e. before and after the network structure change) and

$$p_{ij} = Pr\{\xi_1 = e_j | \xi_0 = e_i\},\ i = \overline{1,n},\ j = \overline{1, n+1}.$$

Lemma 2. Let $\xi_0 = X_0(\theta)$ be the set of solutions to optimization problem (4) and the random value θ has a continuous distribution on $[0, +\infty)$, then the random set $X_0(\theta)$ consist of a single point with probability 1.

This statement follows from Property 2 and the continuity of the distribution of θ.

The set of solutions to optimization problem (9)–(10) $\xi_1 = X_1(\theta)$ consists of a single point with probability 1 too, if $\theta = \theta(\omega)$ is the random continuously distributed valuable.

Theorem 1. Let $\xi_0 = X_0(\theta)$ be the solution to optimization problem (4) and $\xi_1 = X_1(\theta)$ is the solution to optimization problem (9)–(10) with the same random value $\theta = \theta(\omega)$.

If the random value θ has a continuous distribution on $[0, +\infty)$, then the sets $X_0(\theta)$, $X_1(\theta)$ consist of a single point with probability 1 and transition probabilities satisfy the following conditions

$$p_{ik} = 0 \text{ for all } i = \overline{1, n}, \quad k = \overline{1, n}, \quad k \neq i. \tag{11}$$

Thus, the matrix of transition probabilities has the form

$$P = \{p_{ij}\} = \begin{pmatrix} 1 - p_{1,n+1} & 0 & \dots & 0 & p_{1,n+1} \\ 0 & 1 - p_{2,n+1} & \dots & 0 & p_{2,n+1} \\ 0 & 0 & 1 - p_{3,n+1} & \dots & p_{3,n+1} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 - p_{n,n+1} & p_{n,n+1} \end{pmatrix}. \tag{12}$$

Theorem 1 follows from Lemma 1.

In the considered model a passenger who prefers some kind of transport before the changing of the network structure either retains his choice or changes it to the newly introduced one, if it is preferable in terms of the criteria. In other words, if the distribution of the random parameter θ is the same at both stages (i.e. before and after the network structure change), then $p_{ik} = 0$ for $i, k = \overline{1, n}$, $k \neq i$.

The proposed model of passenger preferences is more theoretical at this stage of research. Actual data has a more complex form as a rule. For example, travel by the same type of transport has a different cost depending on the level of comfort, season, etc. Moreover the travel time is not defined uniquely and often depends on random effects. Thus, the proposed model requires verification and possible complications.

IDENTIFICATION OF PASSENGER PREFERENCE FUNCTION BY DISTRIBUTION OF PASSENGER CHOICES

Let's move from the variable $\tau \in [0, +\infty)$ to the

$$\mu = \mu(\tau) = \frac{\tau}{1 + \tau} \in [0, 1]$$

and consider the problem

$$\begin{aligned} &x_1 + \ldots + x_n = 1, \quad x_i \in \{0; 1\}, \quad i = \overline{1, n}, \\ &f_\mu(X) = (1 - \mu)C(X) + \mu T(X) \to \min_X . \end{aligned} \tag{13}$$

Let $U(\mu)$ denote the solution to this problem for a fixed μ. It is easy to see that problem (4) is equivalent to problem (13) in the sense that $X(\tau)$ is the solution to problem (4) if and only if $U(\mu(\tau))$ is the solution to problem (13).

Property 3. The alternative e_k is the solution to problem (13) if and only if $e_k \in E_0$ and

$$L(k) \leq \mu \leq R(k), \tag{14}$$

where

$$L(k) = \max_{t_j > t_k, j = \overline{1,n}} \left\{ 0, \frac{|c_k - c_j|}{|t_j - t_k| + |c_k - c_j|} \right\}, \quad R(k) = \min_{t_j < t_k, j = \overline{1,n}} \left\{ \frac{|c_k - c_j|}{|c_k - c_j| + |t_j - t_k|}, 1 \right\}. \tag{15}$$

The statement follows from Property 2 and the equivalence of problems (13) and (4), because $\mu(\tau)$ is a monotone increasing function.

Assumption 1. Let the alternatives $\{e_1, \ldots, e_n\}$ satisfy the following conditions:

1. $e_k \in E_0$ for all $k = \overline{1, n}$. If this is not so, we remove not Pareto optimal solutions and renumber the rest.

2. The pairs (t_k, c_k), $k = \overline{1,n}$ are numbered in descending order of time t_k (in increasing order of cost c_k).
3. The inequality

$$L(k) < R(k), \tag{16}$$

where $L(k)$, R_k are defined by equalities (8), hold for all $k = \overline{1,n}$. If this is not true for some k, we remove (t_k, c_k) and renumber the rest.

Thus, for such pairs (t_k, c_k), $k = \overline{1,n}$, there exist numbers μ_0, μ_1, ..., μ_k,..., μ_n, that

$$\begin{gathered} 0 = \mu_0 < \mu_1 < \ldots < \mu_n = 1, \\ \mu_k = R(k-1) = L(k), \quad k = \overline{1,n}. \end{gathered} \tag{17}$$

Let us consider problems of distribution identification for the random variable $\mu = \mu(\omega)$ via data on the cost, the time and the passenger preference.

Problem 1. Let the following data be known:

- trip times $T = (t_1, \ldots, t_n)$ and trip costs $C = (c_1, \ldots, c_n)$ for n possible alternative transport modes $\{e_1, \ldots, e_n\}$.
- the data on passenger preference $\{\hat{p}_j, j = \overline{1,n}\}$, where $\hat{p}_j$ is a share of passengers choosing the mode e_j.

These shares (empirical probabilities of the alternatives) determine an empirical distribution of the solution to problem (13) with the random parameter $\mu = \mu(\omega)$. The problem is to find the distribution $f_\mu(x)$ of the random variable μ such that the solution to problem (13) has a probability distribution close to the empirical distribution in some sense.

It is clear that the existence and uniqueness of such distribution depend on how we understand the closeness of the empirical and theoretical distributions and in which class $\mathcal{F}$ we are looking for the distribution $f_\mu(x)$.

Denote theoretical probabilities of the alternatives by

$$p_k = p_k(f_\mu) = Pr\{U(\mu) = e_k\}, \quad k = \overline{1,n},$$

where $U(\mu)$ is the solution to problem (13), the random variable μ has a density $f_\mu(x)$. Here the closeness of the theoretical and the empirical probabilities may be considered in the sense of minimum the following function

$$\Phi(f_\mu) = \sum_{k=1}^{n} \frac{(\hat{p}_k - p_k)^2}{p_k} \to \min_{f_\mu \in \mathcal{F}}. \tag{18}$$

The theoretical distribution allows us to forecast the passenger traffic for a new mode of transport such as high-speed railway. Other methods of passenger traffic forecasting were considered in [5, 6].

If there is a sufficiently large amount of homogeneous data for different transport directions, but for the same period and in the same country, then it is possible to formulate a more general problem of distribution identification for the random variable μ.

Problem 2. Let passengers use M different transport direction and n_m possible transport modes in m-th direction. Let the following data be known:

- trip times $T_m = (t_{m1}, \ldots, t_{mn_m})$ and trip costs $C_m = (c_{m1}, \ldots, c_{mn_m})$ for n_m possible alternative transport modes for m-th direction, $m = \overline{1,M}$.
- the data on passenger preferences $\hat{P}_m = \{\hat{p}_j^{(m)}, j = \overline{1,n_m}\}$, where $\hat{p}_j^{(m)}$ is a share of passengers choosing transport mode $e_j^{(m)}$ in m-th direction.

Denote by

$$A_m = \{e_1^{(m)}, \ldots, e_{n_m}^{(m)}\},$$

a set of Pareto-optimal alternatives in m-th direction, $m = \overline{1,M}$. Let for every $m = \overline{1,M}$ alternatives $\{e_1^{(m)}, \ldots, e_{n_m}^{(m)}\}$ satisfy Assumption 1.

A set of solutions to the problem (13) for a fixed μ and $T = T_m$, $C = C_m$ is denoted by $U_m(\mu)$. Each pair $\{T_m, C_m\}$ corresponds to a set of the numbers $\{\mu_0^{(m)}, \mu_1^{(m)}, ..., \mu_{n_m}^{(m)}\}$ satisfying (17).

We propose that all empirical probabilities (the shares of passengers choosing j-th transport mode in m-th direction) correspond the same random valuable $\mu = \mu(\omega)$, i.e. passengers in all direction have the same distribution of the time cost.

Let denote the density function of μ by $f_\mu(x) \in \mathcal{F}$ and

$$P_m = \{p_k^{(m)}\}, \;\; p_k^{(m)} = Pr\{U_m(\mu) = e_k\}, \;\; k = \overline{1, n_m}, \; m = \overline{1, M}.$$

We get

$$p_k^{(m)} = \int_{\mu_{k-1}^{(m)}}^{\mu_k^{(m)}} f_\mu(x)dx.$$

The problem is to find the theoretical distribution $f_\mu(x) \in \mathcal{F}$ such that

$$\Phi(P_1, P_2, ..., P_M, \hat{P}_1, \hat{P}_2, ..., \hat{P}_M) \to \min_{f_\mu \in \mathcal{F}},$$

where $\Phi(P_1, P_2, ..., P_M, \hat{P}_1, \hat{P}_2, ..., \hat{P}_M)$ is a criterion of closeness between sets of the vectors $P_1, P_2, ..., P_M$ and $\hat{P}_1, \hat{P}_2, ..., \hat{P}_M$.

The numerical solution to the Problem 2 allows to obtain the theoretical distribution of μ.

CONCLUSION

Properties of the probabilistic model for describing passenger preferences are studied. Choosing a route by a randomly selected passenger is considered as the problem of minimizing of the generalized cost of transportation, which is a linear combination two criteria and depends on a random value of the passenger preferences. A mathematical model with random weighting coefficients depending on a randomly selected consumer is used for describing consumer preferences and predicting the correspondence matrix. The problems of estimating the random weights distribution are formulated. The proposed approach may be extended to a wide class of problems of choosing the optimal route, predicting the correspondence matrix and passenger traffic forecasting.

ACKNOWLEDGMENTS

The investigation was supported by the Russian Foundation for Basic Research, project no. 17-08-01123-a.

REFERENCES

[1] G. Timofeeva and A. Martynenko, Analysis of Transport Network Development via Probabilistic Modelling. In: *Stability and Oscillations of Nonlinear Control Systems (STAB-2018): 14th Intern. Conf. (Pyatnitskiy's Conference)*, (ICS RAS, Moscow, 2018), 2p. IEEE Xplore Digital Library, Proceedings.

[2] G. Timofeeva, A. Martynenko and D. Zavalishchin, Probabilistic modeling of passengers and carriers preferences via bicriterial approach. In: 17th IFAC Workshop on Control Applications of Optimization (CAO 2018), IFAC-PapersOnLine, **51**(32), 496–498 (2018).

[3] G. de Rus and V. Inglada, Cost-benefit analysis of the high-speed train in Spain. The Annals of Regional Science **3**, 175–188 (1997).

[4] G. de Rus, *Economic evaluation of the High Speed Rail* (Elanders Sverige AB, Stockholm, 2012)

[5] M. Burjesson, Forecasting demand for high speed rail. Transportation Research Part A: Policy and Practice **70**, 81–92 (2014).

[6] S. Wei, Exploring the potential of open big data from ticketing websites to characterize travel patterns within the Chinese high-speed rail system. PloS one **6**, 17–23 (2017).